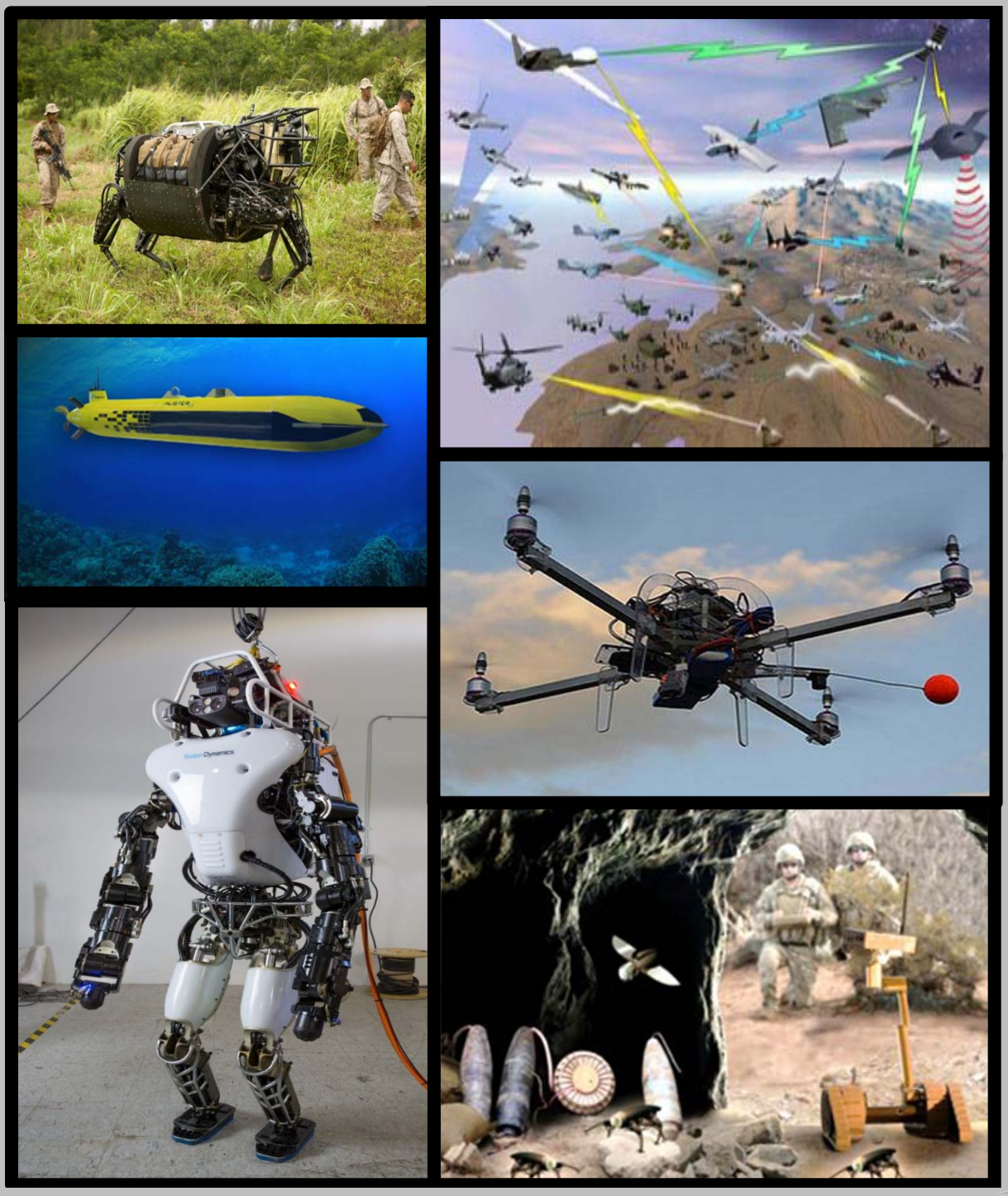
Department of Defense
Research & Engineering

Autonomy Community of Interest (COI)
Test and Evaluation, Verification and Validation (TEVV) Working Group

Technology Investment Strategy
2015-2018



Office of the Assistant Secretary of Defense
For Research & Engineering
May 2015

THE ASSISTANT SECRETARY OF DEFENSE
3030 DEFENSE PENTAGON
WASHINGTON, DC 20301-3030

RESEARCH
AND ENGINEERING

JUN 12 2015

MEMORANDUM FOR THE S&T DEPUTIES COUNCIL

SUBJECT: Autonomy Test and Evaluation, Verification and Validation Technology Investment Strategy 2015-2018

In a April 19, 2011 memo, the Secretary of Defense designated Autonomy as one of the seven priority S&T investment areas in the FY13-17 Program Objective Memorandum. In response, the Assistant Secretary of Defense, Research and Engineering, (ASD(R&E)), set up the Autonomy COI and identified four challenge areas and designated a working group (WG) for each:

- Human/Autonomous Systems Interaction and Collaboration
- Scalable Teaming of Multiple Autonomous Systems
- Machine Reasoning, Perception, and Intelligence
- Test & Evaluation and Verification & Validation (ATEVV)

The ATEVV WG, Co-Chaired by the Defense Threat Reduction Agency (DTRA) and the Air Force Research Laboratory (AFRL) and supported by representatives from DoD's R&D and T&E organizations produced the DoD Autonomy COI Test and Evaluation, Verification and Validation Technology Investment Strategy 2015-2018. The purpose of this strategy is to align DoD Research in TEVV of Autonomy around the following goals:

- **Goal 1 – Methods & Tools Assisting in Requirements Development and Analysis:** Precise, structured standards to automate requirement evaluation for testability, traceability, and de-confliction.
- **Goal 2 – Evidence-Based Design and Implementation:** Assurance of appropriate decisions with traceable evidence at every level of design; reducing current T&E burden.
- **Goal 3 – Cumulative Evidence through RDT&E, DT, & OT:** Progressive sequential modeling, simulation, test and evaluation.
- **Goal 4 – Run Time Behavior Prediction and Recovery:** Real time monitoring, just-in-time prediction, and mitigation of undesired decisions and behaviors.
- **Goal 5 – Assurance Arguments for Autonomous Systems:** Reusable arguments based on previous evidence "building blocks."

Please contact Mr. Matthew Clark at, matthew.clark.20@us.af.mil if you would like any additional information on this strategy.

Alan R. Shaffer
Principal Deputy

cc:
DR. JAGADESSH PAMULAPATI
CAPT. ROBERT PALISIN
COL. CHARLES ORMSBY
DR. JONATHAN A. BORNSTEIN

Table of Contents

Background

In the past decade, unmanned systems have significantly impacted warfare worldwide. They have extended human reach via persistent capabilities, offering warfighters more options to access sensitive and hazardous environments at a speed and scale beyond manned capability. However, current unmanned systems operate with minimal autonomy. To meet warfighter needs and increase military utility, future unmanned systems must have increased autonomy to reduce the cognitive load, improve performance through increased operational speed, and increase performance in denied environments. Recognizing the importance of autonomy, the Secretary of Defense in a 19 April 2011 memo designated autonomy as one of seven priority S&T investment areas in the FY13-17 Program Objective Memorandum. In response to the memo, ASD(R&E) set up Priority Steering Councils (later designated as Communities of Interest (COIs)) for each of these seven areas. The Autonomy COI conducted a series of meetings, identified the following four non-exhaustive technical challenge areas, and designated a working group for each area:

- Human/Autonomous Systems Interaction and Collaboration
- Scalable Teaming of Multiple Autonomous Systems
- Machine Reasoning, Perception, and Intelligence
- Test & Evaluation and Verification & Validation

For the purposes of this document, it is necessary to adhere to a definition of *automation* and *autonomy* in order to address the challenges facing Test & Evaluation and Verification & Validation (TEVV) of such systems. A great debate continues to attempt to define these two terms and their relationship with each other.[1] However, it is in our interest to provide definitions that, at the least, are suitable to address systems being developed by the other three Autonomy COI working groups. Therefore, the Autonomy COI TEVV working group has elected to refer to the following definitions[2]:

> Automation: The system functions with no/little human operator involvement; however, the system performance is limited to the specific actions it has been designed to do. Typically these are well-defined tasks that have predetermined responses (i.e., simple rule-based responses).
>
> Autonomy: The system has a set of intelligence-based capabilities that allows it to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. Autonomous systems have a degree of self-government and self-directed behavior (with the human's proxy for decisions).

TEVV is a critical element for building high assurance of autonomy. Military system developers seek to add increased autonomy to unmanned systems to provide fully functional, self-governing teammates that enhance human warrior operational capability across echelons. This progression requires a significant increase in the trust of autonomous, self-governing systems. This issue with trust is not uncommon; industries that incorporate automation and autonomy with cyber-physical systems struggle with acceptance of new technology, both by users of the technology and with respect to methods used to formally verify performance and safety[3]. It should be noted that although the majority of this document will refer to TEVV of autonomous **systems**, the most difficult and challenging component of these systems is the intelligent, learning, and adaptive **software** embedded within them. Therefore, many of the challenges, gaps, and strategic endeavors listed may appear to be software centric. However, the working group recognizes that, especially when referring to system validation, considerations must be made on how to evaluate the autonomous software agents within the context of the larger cyber-physical systems in which they are employed. The working group also did not specifically include autonomous or semi-autonomous cyberspace systems for cyberspace operations, but techniques or approaches that are articulated in this document could apply to cyberspace systems.

The notion that autonomous systems can be fully tested is becoming increasingly infeasible as higher levels of self-governing systems become a reality. As these systems react to more environmental stimuli and have larger decision

---

[1] Robin Murphy, and James Shields, *Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems*, Technical Report CD 1172, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, Washington, DC (July 2012), http://www.dtic.mil/get-tr-doc/pdf?AD=ADA566864.

[2] Major General Masiello, *Air Force Research Laboratory Autonomy Science and Technology Strategy*, Strategy Report, Air Force Research Laboratory, Wright-Patterson AFB, Ohio (December 2013), http://www.defenseinnovationmarketplace.mil/resources/AFRL_Autonomy_Strategy_DistroA.PDF.

[3] J. Rupp and A. King, *Autonomous Driving - A Practical Roadmap,* SAE Technical Paper (Jan 2010), http://papers.sae.org/2010-01-2335/.

spaces, the standard practice of testing all possible states and all ranges of inputs to the system becomes an unachievable goal. Existing TEVV methods are, by themselves, insufficient for TEVV of autonomous systems; therefore a fundamental change is needed in how we validate and verify these systems. For example, agent- and model-based design and verification techniques and simulations are gaining ground as methods to guarantee assure safety in multi-agent systems in industry; however, acceptance of this paradigm shift toward simulation-based validation and verification has yet to replace many physical tests in military programs of record.[4] The challenge of TEVV of autonomous systems was highlighted in the 2010 Air Force Technology Horizon report, which states that, "It is possible to develop systems having high levels of autonomy, but it is the lack of suitable V&V methods that prevents all but relatively low levels of autonomy from being certified for use."[5] Additionally, the Defense Science Board released a report on the Role of Autonomy in the Department of Defense (DoD) Systems, recommending "The Under Secretary of Defense for Acquisition, Technology and Logistics (USD (AT&L)) should create developmental and operational test and evaluation (T&E) techniques that focus on the unique challenges of autonomy (to include developing operational training techniques that explicitly build trust in autonomous systems). DoD needs new technology to assist the test community with certifying systems at the end of development – a situation that has not yet happened because currently fielded autonomy technologies have by-passed the formal test process due to the pressing demands of the recent conflicts.[1]"

In 2014, the Autonomy COI TEVV working group became fully functional, with joint representatives within the Department of Defense. These representatives joined the working group in an effort to coordinate research in the V&V of Autonomy. Currently, the Autonomy COI TEVV working group consists of representatives from the Defense Threat Reduction Agency (DTRA), Air Force Research Laboratory (AFRL), Naval Research Laboratory (NRL), Office of Naval Research (ONR), Army Research Laboratory (ARL), US Army Aviation and Missile Research Development and Engineering Center (AMRDEC), Test Resource Management Center (TRMC), Aberdeen Test Center (ATC), Marine Corps Intelligence Activity (MCIA), the Office of Naval Intelligence (ONI), Naval Sea Systems Command (NAVSEA), and Joint Ground Robotics Enterprise (JGRE). Additionally, we have collaborators from MITRE, Georgia Tech Research Institute (GTRI), Applied Research Associates (ARA), The Analytic Sciences Corporation (TASC), Johns Hopkins University / Applied Physics Laboratory (JHU/APL), and the LinQuest Corporation.

[4] Josie Hunter, "A Synergistic and Extensible Framework for Multi-Agent System Verification," *12th International Conference on Autonomous Agents and Multiagent Systems*, May 2013, http://www.aamas-conference.org/Proceedings/aamas2013/docs/p869.pdf.

[5] W. J. A. Dahm, *Technology Horizons a Vision for Air Force Science & Technology During 2010-2030*, USAF HQ (2010).

Purpose

Our purpose is to lay the groundwork for a technology roadmap, identifying research objectives to further the state of the art in TEVV of Autonomous Systems. This document strives to accomplish this by taking the following actions:

- Highlighting the challenges to TEVV of Autonomous Systems.
- Identifying the Goals and Methodologies needed to address these challenges.
- Providing a framework to align Department of Defense Projects / Products.

The Autonomy COI TEVV investment strategy endeavors not only to codify the challenges and gaps of TEVV for autonomy, but more importantly, to identify a way forward.

Current Challenges to Autonomy TEVV

In FY13, the Autonomy COI TEVV (ATEVV) working group, in coordination with the AFRL Autonomy TEVV group, had several workshops to identify the enduring challenges of both what makes the test and evaluation of autonomy difficult and what roadblocks or gaps exist in the current systems engineering and T&E infrastructure. Those studies identified the following enduring problems for future TEVV of Autonomous Systems:

- **ATEVV Challenge 1 - State-Space Explosion:**
  Autonomous systems are characteristically adaptive, intelligent, and/or may incorporate learning. For this reason, the algorithmic decision space is either non-deterministic, i.e. the output cannot be predicted due to multiple possible outcomes for each input, or is intractably complex. Because of its size, this space cannot be exhaustively searched, examined, or tested; it grows exponentially as all known conditions, factors, and interactions expand. Therefore there are currently no established metrics to determine various aspects of success or comparison to a baseline state enumerated.

- **ATEVV Challenge 2 - Unpredictable Environments:**
  The power of autonomous agents is the ability to perform in unknown, untested environmental conditions. Examples of environmental “stimuli” include actors capable of making their own decisions in response to autonomous system actions; producing a cognitive feedback loop that explodes the state space. Additionally, autonomous decisions are not necessarily memoryless and the state space is not just the intractably complex in the current situation, but also in the multiplicity of situations over time. Currently fielded systems have very limited robustness to dynamic / changing environmental conditions. Adaptive autonomous algorithms have the potential to overcome current automated system brittleness in future dynamic, complex, and/or contested environments. However, this performance increase comes with the price of assuring correct behavior in a countless number of environmental conditions. This exacerbates the state-space explosion problem.

- **ATEVV Challenge 3 - Emergent Behavior:**
  Interactions between systems and system factors may induce unintended consequences. With complex, adaptive systems, how can all interactions between systems are captured sufficiently to understand all intended and unintended consequences? How can autonomous design approaches identify or constrain potentially harmful emergent behavior both at design time and at run time? What limitations are there with the current Design of Experiments approach to test vector generation when considering adaptive decision-making in both discrete decision logic and continuous variables in an unpredictable environment? Since emergent behavior can be produced by interactions between small, seemingly insignificant factors how can we provide test environments or test oracles that are of sufficient fidelity to examine and capture emergent behavior (in M&S, T&E, and continuous operations or run time testing)?

- **ATEVV Challenge 4 - Human-Machine Communication:**
  Handoff, communication, and interplay between operator and autonomy become a critical component to the trust and effectiveness of an autonomous system. Current certification processes eliminate the need for “trust” through exhaustive Modeling and Simulation (M&S) and T&E to exercise all possible operational vignettes. When this is not possible at design time, how can trust in the system be ensured, what factors need to be addressed, and how can transparency and human-machine system requirements for the autonomy be defined?

Changes Needed in the Current V&V Paradigm

Current modeling, simulation, test and evaluation methods, though effective for countless currently fielded systems, become a bottleneck are infeasible when attempting to field systems that include higher levels of autonomy. The working group highlighted a larger systemic problem in fielding autonomous systems. Namely, the difficulty of TEVV is inversely proportional to the amount of early design time V&V performed. Yet relatively little attention is paid to V&V of requirements, architectures, and early design models. Therefore, the ATEVV working group highlighted the following challenges, not merely in test and evaluation but in a broader systems engineering context:

- **ATEVV Gap 1 – Lack of Verifiable Autonomous System Requirements:**
  Currently, there is a lack of common, clear, and consistent requirements for systems that include autonomous requirements, especially with respect to environmental assumptions, Concept of Operation (CONOPS), interoperability, and communication. There is also a lack of clearly defined Measures of Effectiveness (MOEs), performance measures, and other metrics. For instance, requirements and metrics are needed for:
  - The detection of cases in which autonomous decisions violate system deviations in autonomous requirements from the higher level system requirements and MOPs that describe conditions under which specifications are, or are not, violated.
  - Effectiveness of human-automation interaction, e.g. with respect to human workload and situational awareness
  - System performance relative to that of a human operator
  - Demonstrable confidence level / acceptable risk of autonomous systems
  - Interoperability and security
  - Performance of intra-agent, intra-system collaboration (e.g., 10+ agents)

  Furthermore, there are deficiencies in ensuring the traceability of requirements to implementation (e.g., manufacturing or compiling) both manually and automatically. Automatic requirements extraction and validation is needed for future learning / adaptive systems. Finally, current autonomous systems requirements are not written or analyzed to ensure that they are verifiable.

- **ATEVV Gap 2 – Lack of Modeling, Design, and Interface Standards:**
  Currently, no standardized modeling frameworks exist for autonomous systems that span the whole system lifecycle (R&D through T&E). Therefore, a gap exists in traceability between capabilities implemented in conventional systems as well as adaptive, nonlinear, stochastic, and/or learning systems and the requirements they are designed to meet. This results in a need to integrate models that are both heterogeneous and composable in nature and existing at different levels of abstraction, including requirements, architecture models, physics-based models, cognitive models, test range/environment models, etc. Some example approaches that attempt to address this gap include: Future Aviation Common Environment (FACE), model-based engineering, Joint Architecture of Unmanned Systems (JAUS), Navy's Advanced EOD Robotics System (AEODRS), Standardization Agreement (STANAG), RS-JPO Interoperability Profile (IOP), and the UAS Control Segment Architecture (UCS). Additionally, there is a need to formalize these design ontologies such that future testing can benefit from features they can encode, such as software invariants (operational software "test" points that always hold), behavioral assume/guarantee contracts, and software heartbeats.

- **ATEVV Gap 3 – Lack of Autonomy Test and Evaluation Capabilities:**
  As stated before, there is a current gap in T&E ranges, test beds, and skillsets for handling dynamic learning / adaptive systems.[6] [7] The complexity of these systems results in an inability to test under all known conditions, difficulties in objectively measuring risk, and an ever-increasing cost of rework / redesign due to errors found in late developmental and operational testing. Furthermore, the lack of formalized requirements and system models makes test-based instrumentation of model abstractions increasingly difficult. This limits design-for-test capabilities, including tests to evaluate human-autonomy interactions.

- **ATEVV Gap 4 – Lack of Human Operator Reliance to Compensate for Brittleness:**
  Currently, the burden of decision making under uncertainty is placed solely on human operators. Certification, acceptance, and risk mitigation often assume the human operator can compensate for the brittleness currently found in manned, remotely piloted, or tele-operated systems. However, as systems move from relatively predictable automated behaviors to more unpredictable and complex autonomous behaviors, and as autonomous systems operate in denied environments in which they interact with human intermittently, it will become increasingly difficult for human operators to understand and respond appropriately to decisions made by the system. Thus, V&V of autonomous software should also take into account factors relating to human-machine interfaces, human performance characteristics, requirements for human operator training, etc. As the National Research Council states in a report on autonomy in civilian aviation, "In twentieth century aviation, system-level "intelligence" resided almost exclusively with human operators. The segregated assessment of the aircraft and crew for flight readiness was acceptable, because the standards were for the most part independent. That is not the case with IA *(increasingly autonomous)* systems, where the intelligence is likely to be divided between the human and the system. Even if a V&V method were to prove useful for assessing intelligent software, the assessment of the total system, including the human operators and their interactions with the IA system, requires another approach." [7]

[6] Robin Murphy, and James Shields, *Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems*, Technical Report CD 1172, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, Washington, DC (July 2012), http://www.dtic.mil/get-tr-doc/pdf?AD=ADA566864.

[7] National Research Council, *Autonomy Research for Civil Aviation: Toward a New Era of Flight*, Washington, DC: The National Academies Press, 2014, http://www.nap.edu/catalog/18815/autonomy-research-for-civil-aviation-toward-a-new-era-of.

- **ATEVV Gap 5 – Lack of Run Time V&V during Deployed Autonomy Operations:**
  As stated earlier, current automated systems rely on human oversight to guarantee safe and correct operation, with the human operator acting as the ultimate monitor, kill switch, and recovery mechanism for brittle automation. [8] However, as systems incorporate higher levels of autonomy, it will no longer be feasible or safe to rely solely on human operators for system monitoring and recovery. Therefore, "sandboxing," bounding, or encapsulation of learning / adaptive systems must be developed and supported in the V&V process so that higher levels of autonomy can be deployed in operational environments without full, exhaustive testing. [9] [10]

- **ATEVV Gap 6 – Lack of Evidence Re-use for V&V:**
  Results from TEVV do not, by themselves, accept operational risk, imply certification, or give authority to operate. However, TEVV results provide the collected body of evidence that is presented to a certification board, and ultimately the milestone decision authority (MDA), to determine an acceptable level of safety, security, performance, and risk for that specific platform. Certification standards aid programs in providing the appropriate artifacts generated through T&E activities. However, current assurances (arguments of safety and security that a system falls within an acceptable level of risk) are predominately manual, subjective, and often not reusable. [11] [12] Additionally, current regulations "imply" a formal argument of system assurance based on tradition and past failure conditions of "similar" systems, which may not exist for groundbreaking autonomous systems. This implication could impede rapid deployment of future autonomous systems by causing unsustainable T&E requirements to be generated from ill-defined and duplicative arguments of acceptable risk. [13]

[8] David McNally, "Army Robotics Researchers Look Far into the Future," *www.army.mil* 14 Nov. 2014, http://www.army.mil/article/137837/Army_robotics_researchers_look_far_into_the_future/

[9] Kerstin Eder, et al., "Assurance Using Models at Runtime for Self-Adaptive Software Systems," *Lecture Notes in Computer Science*, Volume 8378, 2014, pp 101-136.

[10] Matthew Clark, et al., *A study on Run Time Assurance for Complex Cyber Physical Systems*, Air Force Research Laboratory, Aerospace Systems Directorate, Wright-Patterson AFB OH, 2013.

[11] John Rushby, *Modular Certification*, National Aeronautics and Space Administration, Langley Research Center, 2002.

[12] Ewen Denney, and Ganesh Pai, "Evidence Arguments for Using Formal Methods in Software Certification," *IEEE International Workshop on Software Certification (WoSoCer 2013)*, November 2013.

[13] C. Michael Holloway, "Making the Implicit Explicit: Towards an Assurance Case for DO-178C," *31st International System Safety Conference*, 12-16 August 2013, Boston, MA.

Vision

The future state envisions an “autonomous agent” no longer restricted by the inability to be certified as trustworthy at an acceptable level of operation and risk. Acknowledging that the “autonomous agent” can take many forms, e.g. as a rational decision maker for manned / unmanned aircraft, cyber systems, satellites, or weapons, the vision depicts a future where **alternate evidence** of verification and validation can be generated through additional techniques in addition to current M&S and T&E methods. The results from these methods can be recorded in a **modular** fashion, enabling **compositional** verification of autonomous subcomponents at appropriate levels of abstraction, thereby reducing the system-level V&V challenge. Additionally, similar to case law, well-defined and iteratively developed autonomous agents will be able to establish **a precedent** through past performance and use of “training” as a method of certification. Finally, development of autonomous agents will be **iterative, continuous**, and **evolutionary**, reducing the software development cycle burden.

Toward this end, AFRL researchers have proposed that the classic “V” used to describe the software development process be modified to incorporate verification activities throughout the development cycle (see Figures 1 & 2). Figure 2 shows a conceptual Autonomy Community of Interest TEVV Process Model that integrates development and V&V, with V&V activities occur during and between each major development activity. With this process model, we endeavor to “flatten” the systems engineering “V” through the incremental and compositional assurances (or arguments) of safety, security, performance, and risk.

For future highly autonomous systems, verification and validation activities, including where, when, and to what degree V&V happens, must be significantly emphasized throughout the complete acquisition process. Testing and Evaluation is and must be a significant part of overall autonomous system V&V activities. Testing and evaluating future highly autonomous systems will require an increased emphasis on setting verifiable requirements, developing system models traceable to requirements to guide design activities, and verifying and validating emerging subsystems and products throughout the development process. The final, traditional Development Test (DT) and Operational Test (OT) activities must become a final verification of the complete body of evidence leading to and supporting the documentation of the safety, effectiveness, suitability, and survivability of the system. Essentially, the addition of autonomy will require that much of the effort traditionally reserved for final DT and OT of a new system must be shifted to the left, with the majority of the T&E activities taking place before the completed system is assembled at test ranges for final system level DT and OT.

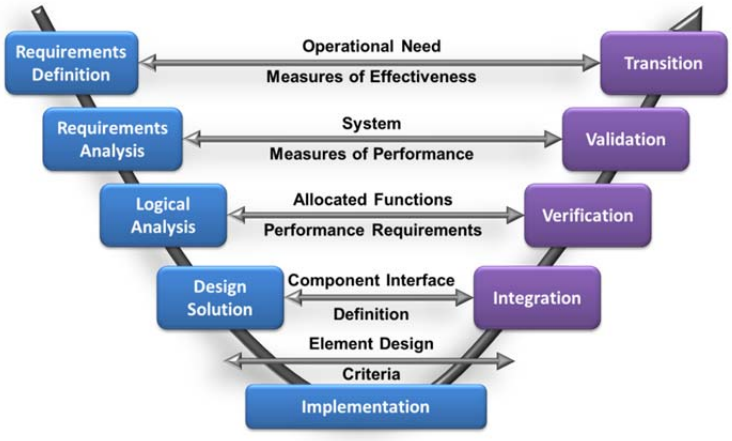


Figure 1 Classic “V” Development Cycle

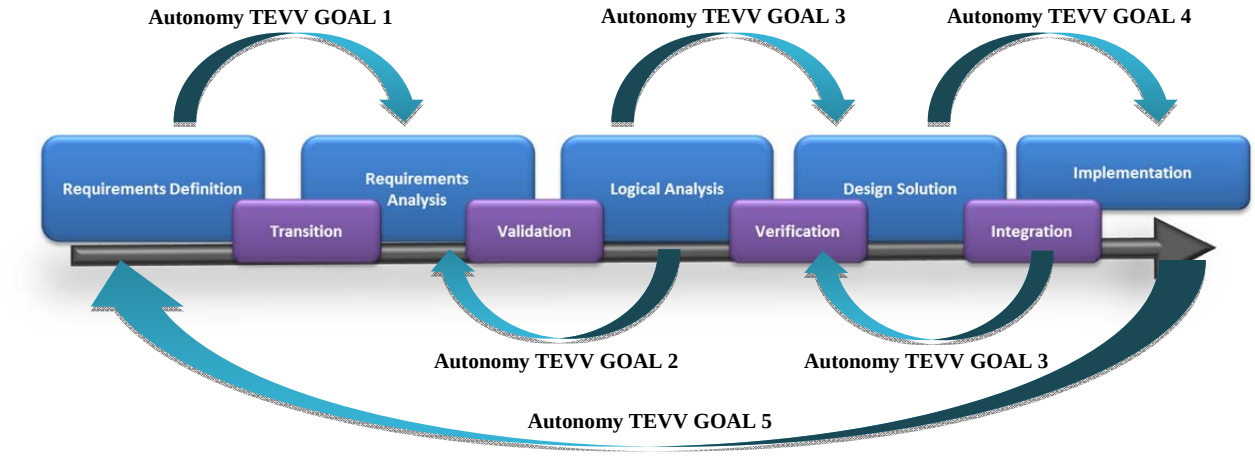


Figure 2: Autonomy TEVV Process Model, Integrated with the traditional Systems Engineering "V"

Autonomy TEVV Goals

The following technical goals provide multiple, additive methods for V&V of autonomous systems at each point in the systems engineering lifecycle. The purpose of this document and the ATEVV working group is to align DoD Research and Development programs around these goals in an effort to fill ATEVV gaps and thereby overcome the current ATEVV challenges.

TEVV Goal 1: Methods & Tools Assisting in Requirements Development and Analysis

Precise, structured standards to automate requirement evaluation for testability, traceability, and de-confliction

Department of Defense Instruction 5000.02 states that all "acquisition programs respond to validated capability requirements" throughout the acquisition process.[14] At each milestone decision authority, capability requirements are revised, matured, and used as the primary basis for verification. Improved tools and methods to help the systems engineering community better articulate, formalize, and validate autonomous requirements are therefore paramount to the success of system verification and validation. This goal focuses on ***increasing the fidelity and correctness of autonomous system requirements by developing methods and tools to enable the generation of requirements that are, where possible, mathematically expressible, analyzable, and automatically traceable to different levels (or abstractions) of autonomous system design.*** For instance, tools that map constrained natural language to formalized, mathematically rigorous semantics allow subject matter experts who develop requirements to be explicit and to clearly define assumptions. Additionally, such formalisms provide high-level operational assumptions and interoperability guarantees that can be analyzed early in the design phase. Equal vigor must address not only functional requirements of the autonomous system but also equivalent models of the environment with which it will interact, utilizing constructive and/or hybrid approaches that compose requirements on the basis of well-formed mathematically rigorous statements and statistical requirements. Finally, requirements aren't complete without an understanding of how they will be tested. Formalized requirements enable automatic test generation and traceability to low-level designs, but note that TEVV representatives must be involved early in the requirements development process. Specific requirements and requirement templates must be constructed that articulate how autonomous vehicles need to perform in unknown and untested environmental conditions that can induce unintended consequences.

[14] *DoD Instruction 5000.02: Operation of the Defense Acquisition System*, January 7, 2015.

TEVV Goal 2: Evidence-Based Design and Implementation

Assurance of appropriate decisions with traceable evidence at every level of design to reduce the current T&E burden

The increasing cost of test is continuing to drive acceptance of higher risk. The prohibitive cost of testing for adaptive, learning, and non-deterministic systems places the adoption of autonomous systems in peril. Through "design for certification" – the increased use of formal, mathematically rigorous or standardized process-based design starting at the beginning of the design process –substantial gains in reduced planning time, test resources, and test cost will be realized. ***Methods and tools need to be developed at every level of design from architecture definition to modeling abstractions to software generation / hardware fabrication, enabling the compositional verification of the progressive design process, thereby increasing test and evaluation efficiency.*** As stated in the AF Technology Horizon report, "Emphasis is on composability via system architectures based on fractionation and redundancy. This involves advancing methods for collaborative control and adaptive autonomous mission planning, as well as V&V of highly adaptable, autonomous control systems."[15] In order to provide assurance for machine intelligence and decision-making in complex, uncertain, and dynamic environments, a paradigm shift must be realized. Similar to the early development of control theory, formal, compositional, and verifiable abstraction-based design must seek to provide proofs about the safety, reliability, performance, risk, and robustness of autonomous systems, creating "correct by construction" designs where possible to reduce the test and evaluation burden.

TEVV Goal 3: Cumulative Evidence through RDT&E, DT, & OT

Progressive sequential modeling, simulation, test and evaluation

M&S and T&E at each Technical Readiness Level (TRL) and product milestone currently provide an invaluable resource not only to verify and validate that a system satisfies the user requirements, but also to aid in technology development and maturation. However, the development of effective methods to record, aggregate, and reuse T&E results remains an elusive and technically challenging problem. As an example, DoD Directive 3000.09 implies that autonomous weapons software, where possible, not be re-written but incrementally developed and verified by sequential, progressive regression testing.[16] It is paramount that products, methods, tools, and capabilities developed in Goal 1 and Goal 2 support the transition of autonomous systems to the DT and OT communities, to better define and, where reasonable, focus and increase the effectiveness of test and evaluation plans. ***Methods must be developed to record, aggregate, leverage, and reuse M&S and T&E results throughout the system's engineering lifecycle; from requirements to model-based designs, to live virtual construction experimentation, to open-range testing.*** This goal endeavors to research the development of standardized data formats and Measures of Performance (MOPs) to encapsulate experimental results performed in early research and development, ultimately reducing the factor space in final operational tests. Additionally, statistics-based design of experiments methods currently lacks the mathematical constructs capable of designing affordable test matrices for non-deterministic autonomous software. Software systems require a risk-mitigation methodology offering the same spirit as Design of Experiments (DOE) while not relying entirely on statistical approaches.

[15] W. J. A Dahm, *Technology Horizons a Vision for Air Force Science & Technology During 2010-2030,* USAF HQ (2010).

[16] *DoD Directive 3000.09: Autonomy in Weapon Systems,* November 21, 2012.

TEVV Goal 4: Run Time Behavior Prediction and Recovery

Real-time monitoring, just-in-time prediction and mitigation of undesired decisions and behaviors

For the most demanding adaptive and non-deterministic systems, we may need even a more dramatic shift. Currently, we attempt to prove systems correct via verification of every possible state PRIOR to fielding of the system, through extensive and costly test and evaluation. ***However, for highly complex autonomous systems, an alternate method leveraging a run-time architecture must be developed that can provably constrain the system to a set of allowable, predictable, and recoverable behaviors, shifting the analysis/test burden to a simpler, more deterministic run-time assurance mechanism.*** On the surface, it may seem disconcerting to allow a system to function without exhaustive testing / analysis. However, this concern is born out of the misguided assumption that current software systems are exhaustively tested and are without errors or defects. Rather, software is run through a series of quality steps, checklists, and verification practices that increase the implicit confidence of safety-critical software. [17] This goal endeavors to provide a structured argument, supported by evidence, justifying that a system is acceptably safe and secure not only through offline tests, but also through reliance on real-time monitoring, prediction, and failsafe recovery. Within this paradigm, formal design approaches such as, but not limited to, "assume-guarantee reasoning" generated from design methods outlined in Goal 2, might provide the offline design considerations and formalisms necessary for articulating the allowable and certifiable behaviors of an advanced, uncertified system and for validating the design of a run-time constraint, prediction, and recovery methods.

TEVV Goal 5: Assurance Arguments for Autonomous Systems

Reusable assurance case based on previous evidence "building blocks"

An *assurance case* can be defined as a structured argument, supported by evidence, intended to justify that a system is acceptably safe and secure. [18] A defensible argument of acceptable risk is required as part of the regulatory process, with a certificate of assurance being granted only when the regulator is satisfied by the argument presented. As previously stated, results from TEVV do not by themselves determine operational risk, imply certification, or give authority to operate. However, TEVV results provide the collected body of evidence that is presented to a certification board, and ultimately the milestone decision authority (MDA), to determine an acceptable level of safety, security, performance, and risk for that specific platform. The assumption is that no one method for verification and validation will be adequate for future autonomous systems. ***Therefore, not only do multiple new TEVV methods need to be employed to enable the fielding of autonomous systems, a new research area needs to be investigated in formally articulating and verifying that the assurance argument itself is valid.*** Research must be done to formalize assurance cases for the purposes of analysis and reuse, providing a comprehensive argument that all requirements have been satisfied, including safety, security, performance, etc. This structured argument-based approach must be developed in coordination with and as an integral part of the Test and Evaluation Plan (TEP) and the Test and Evaluation Master Plan (TEMP), providing a claim of how the V&V activities will endeavor to quantify risks and mitigation strategies to inform risk-acceptance decisions. Additionally, standard autonomy argument templates must be developed, enabling the reuse of explicit arguments of risk, performance, and safety, closely tied to autonomy requirements and TEVV practices which, if performed, provide an acceptable collection of evidence for an autonomous system.

[17] C. Michael Holloway, "Making the Implicit Explicit: Towards an Assurance Case for DO-178C," *31st International System Safety Conference*, 12-16 August 2013, Boston, MA.

[18] Tim Kelly and Rob Weaver, "The Goal Structuring Notation–a Safety Argument Notation," *Proceedings of the Dependable Systems and Networks Workshop on Assurance Cases*, 2004.

Appendix A: Definitions

1. Accreditation

The official certification that a model or simulation and its associated data are acceptable for use for a specific purpose.[19] *(Definition relative to M&S accreditation*) Accreditation is the formal declaration by a neutral third party that the certification program is administered in a way that meets the relevant norms or standards of certification program.[20]

2. Adaptive/Nondeterministic Systems

"Adaptive systems have the ability to modify their behavior in response to their external environment. For aircraft systems, this could include commands from the pilot and inputs from aircraft systems, including sensors that report conditions outside the aircraft. Some of these inputs, such as airspeed, will be stochastic because of sensor noise as well as the complex relationship between atmospheric conditions and sensor readings not fully captured in calibration equations. Adaptive systems learn from their experience, either operational or simulated, so that the response of the system to a given set of inputs varies and, presumably, improves over time. Systems that are nondeterministic may or may not be adaptive. They may be subject to the stochastic influences imposed by their complex internal operational architectures or their external environment, meaning that they will not always respond in precisely the same way even when presented with identical inputs or stimuli. The software that is at the heart of nondeterministic systems is expected to enable improved performance because of its ability to manage and interact with complex "world models" (large and potentially distributed data sets) and execute sophisticated algorithms to perceive, decide, and act in real time." – Autonomy Research for Civil Aviation: Toward a New Era of Flight[21]

3. Assurance Cases

The assurance case provides a means to structure the reasoning that engineers implicitly use to gain confidence that systems will work as expected. It also becomes a key element in the documentation of the system and provides a mapping to more detailed information. The concept of an assurance case has been derived from the safety case, a construct that has been used successfully in Europe for over a decade to document safety for nuclear power plants, transportation systems, automotive systems, and avionics systems. Much like a legal case presented in a courtroom, an assurance case requires arguments linking evidence with claims of conformance to dependability-related requirements.[22] Several certification standards and guidelines in the defense, transportation (aviation, automotive, rail), and healthcare domains now recommend and/or mandate the development of assurance cases for software-intensive systems.[23,24]

4. Assume-Guarantee Reasoning

A form of compositional proof, performed by systematically defining and verifying the pre-conditions (assumptions) and post-conditions (guarantees) that govern the interconnections between all subcomponents within a system. [25] [26] [27]

---

[19] *DoD Instruction 5000.61: DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)*, December 9, 2009.

[20] ISO/IEC 17024, *Conformity Assessment – General Requirements for Bodies Operating Certification of Persons*.

[21] National Research Council, *Autonomy Research for Civil Aviation: Toward a New Era of Flight*, Washington, DC: The National Academies Press, 2014, http://www.nap.edu/catalog/18815/autonomy-research-for-civil-aviation-toward-a-new-era-of.

[22] Software Engineering Institute Carnegie Mellon University, *Assurance Cases*, http://www.sei.cmu.edu/dependability/tools/assurancecase/.

[23] *1st International Workshop on Assurance Cases for Software-Intensive Systems (ASSURE 2013)*.

[24] Richard Hawkins, Ibrahim Habli, and Tim Kelly, "The Principles of Software Safety Assurance," *International System Safety Conference (ISSC)*, 2013, Boston.

[25] K. L. McMillan, "*Circular Compositional Reasoning about Liveness*," Cadence Berkeley Labs, Berkeley, CA 94704, Tech. Rep. 1999-02.

[26] Goran Frehse, Zhi Han, and Bruce Krogh. "Assume-Guarantee Reasoning for Hybrid I/O Autonomata by Over-Approximation of Continuous Interaction," *Decision and Control*, 2004. CDC 43rd IEEE Conference Vol 1. IEEE, 2004.

[27] Michael Huth and Mark Ryan, *"Logic in Computer Science, Modelling and Reasoning about Systems,"* Cambridge University Press, ISBN: 978-0-521-54310-1, 2004

5. Automation

The system functions with no/little human operator involvement; however, the system performance is limited to the specific actions it has been designed to do. Typically these are well-defined tasks that have predetermined responses (i.e., simple rule-based responses).

6. Autonomy

The system has a set of intelligence-based capabilities that allows it to respond to situations that were not pre-programmed or anticipated (i.e., decision-based responses) prior to system deployment. Autonomous systems have a degree of self-government and self-directed behavior (with the human's proxy for decisions).

7. Brittleness/Robustness

Brittleness is the inability of software to cope with errors or new situations during execution. Robustness is the opposite of brittleness.

8. Certification

Certifications provide a formal acknowledgment by an approval authority that a system or program meets specific requirements. Certifications, in many cases, are based on statute or regulations and drive systems engineering (SE) planning (i.e., a program may not be able to test or field the capability without certain certifications).[28]

9. Formal Verification Methods

Formal methods involve the specification of requirements and the design of a system in a formal specification language or ontology that is semantically complete and allows for rigorous analysis. With the requirements and software expressed in the formal language, analysis approaches can broadly be separated into:

Axiomatic approaches: involve reducing the analysis to a mathematical theorem proof and provide mathematically rigorous statements on the ability of the program to achieve the specifications.

Semantic approaches: include model-checking, which utilize exhaustive search through all possible program executions while looking for behavior inconsistent with formally stated requirements.[29]

10. Modeling and Simulation

Modeling is the representation of a physical system by an equation or set of equations (which must be verified as being the correct representation of physical reality). Simulation involves the mathematical solution of those equations (which must be validated against experimental data).[30]

11. Run Time Assurances (RTA)

A structured argument supported by evidence, justifying that a system is acceptably safe and secure not through reliance on offline tests, but through reliance on real time monitoring, prediction, and failsafe recovery.[31]

12. (RTA) Untrusted Component

Software component not certified at the same criticality level to which it resides. Can be considered an unscripted, dynamic, non-deterministic, adaptive, or learning component.

---

[28] "Chapter 4 – Systems Engineering," *Defense Acquisition Guidebook*, 15 May 2013.

[29] R. Scott Erwin, Paul Zetocha, *Spacecraft Autonomy Technology: A Survey*, Air Force Research Laboratory Space Vehicles Directorate, Kirtland AFB, NM 87117, July 26, 2012.

[30] *1490 WG-IEEE Guide: Adoptions of the Project Management Institute (PMI) Standard: A Guide to the Project Management Body of Knowledge (PMBOK Guide) Working Group*, Institute of Electrical and Electronic Engineers (IEEE), 4th ed., 2008, http://standards.ieee.org/develop/wg/software_and_systems_engineering_all.html.

[31] M. Clark, et al., *A Study on Run Time Assurance for Complex Cyber Physical Systems*, Technical Report, WPAFB, 2013, http://www.dtic.mil/docs/citations/ADA585474.

13. (RTA) Recovery System

Set of transition and baseline components providing overall assurance that at any given time, RTA protected system can recover from an untrusted component failure.

14. (RTA) Baseline Component

Software component(s), certified to maintain RTA protected critical functions under specific and limited conditions using deterministic and reliable decision procedures.

15. (RTA) Transition Component

Software component(s) certified to transition the system from any condition at which the untrusted component failed to an operating condition suitable for the baseline controller to engage.

16. (RTA) Monitor & Switch

Certified run time executive that compares the untrusted component behavior with a set of known, acceptable constraints based on the assume-guarantee contracts of each subcomponent interaction and behavior. Monitor determines, based on the violation of a constraint and the time required to recover, when to switch from the untrusted to recovery components.

17. Requirements

Formal system development starts with requirements engineering, which focuses on what a system should do and the constraints under which it must do it. This includes determining the system's functional and non-functional requirements. Functional requirements define what the system will actually do, while non-functional requirements refer to its qualities, such as performance, along with any constraints under which the system must operate. Some non-functional requirements for autonomous systems include:[32]

- Adaptability, which is the system's ability to modify its behavior or structure, which can involve changes in functionality algorithms, systems parameters, and so on. Adaptability requires a model of the system's environment. A key challenge with this requirement is how to measure adaptability.
- Dynamicity refers to the ability to perform a change at run time, such as removing, adding, or exchanging services and components.
- Robustness is the ability to cope with errors or unforeseen situations during execution. Brittleness is essentially the opposite of robustness.
- Resilience is a prerequisite for system agility and safety, and it enables systems to recover from unanticipated disruptions.
- Mobility indicates system re-configurability at both design time and run time and often enables dynamic discovery and usage of new resources.

[32] Emil Vassev, Mike Hinchey, "Autonomy Requirements Engineering," *Computer*, vol.46, no. 8, August 2013, doi:10.1109/MC.2013.267, 82-84.

18. Trust

Trust is not a trait of the system; it is the system status in the mind of human beings based on their perception of and experience with the system. Trust concerns the attitude that a person or technology will help achieve specific goals in a situation characterized by uncertainty and vulnerability.[33] Trust includes measures of trust and both system and operational trust. Research in the field of human-automation interaction (HAI) has shown that trust is a key factor that influences an operator's interaction with an autonomous system. Researchers also found that proper calibration of trust is critical to safe operation of an autonomous system. Too much trust on the system can lead to abuse of automation and conversely too little trust can lead to disuse of automation. In dynamic systems, operators need a control allocation strategy that optimizes performance. Hence, mis-calibrated trust can lead to inefficient operation or even catastrophic failures.[34]

19. Validation

Validation provides objective evidence that the capability provided by the system complies with stakeholder performance requirements, achieving its use in its intended operational environment. Validation answers the question, "Is it the right solution to the problem?" Validation consists of evaluating the operational effectiveness, operational suitability, sustainability, and survivability of the system or system elements under operationally realistic conditions.[35]

20. Verification

Verification provides evidence that the system or system element performs its intended functions and meets all performance requirements listed in the system performance specification and functional and allocated baselines. Verification answers the question, "Did you build the system correctly?"

[33] J.D. Lee and K.A. See, 2004, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, 46(1) (2004), 50-80.

[34] http://robotics.cs.uml.edu/research/trust.php.

[35] "Chapter 4 – Systems Engineering," *Defense Acquisition Guidebook*, 15 May 2013, 164. Ibid, 163.

Appendix B: Acronym List

Acronym	Definition
A2AD	Anti-Access Area Denial
AADL	Architecture Analysis and Design Language
AEODRS	Advanced EOD Robotics System
AFLCMC	Air Force Life Cycle Management Center
AFMC	Air Force Materiel Command
AFOSR	Air Force Office of Scientific Research
AFRL	Air Force Research Laboratory
ALUGS	Appliqué and Large Unmanned Ground Systems
AMRDEC	US Army Aviation and Missile Research Development and Engineering Center
API	Application Programming Interface
ARL	Army Research Laboratory
ARPI	Autonomy Research Pilot Initiative
ATC	Aberdeen Test Center
BAA	Broad Agency Announcement
CONOPS	Concept of Operation
CTEIP	Central Test and Evaluation Investment Program
DASD(DT&E)	Deputy Assistant Secretary of Defense, Developmental Test and Evaluation
DOE	Design of Experiments
DOT&E	Director, Operational Test and Evaluation
DT	Developmental Test
DTRA	Defense Threat Reduction Agency
FACE	Future Aviation Common Environment
FSM	Finite State Machine
IA	Increasingly Autonomous
IOP	Interoperability Profile
JAUS	Joint Architecture of Unmanned Systems
JGRE	Joint Ground Robotics Enterprise
M&S	Modeling and Simulation
MCIA	Marine Corps Intelligence Activity
MDA	Milestone Decision Authority
MIMFA	Machine Intelligence for Mission-Focused Autonomy
MOEs	Measures of Effectiveness
MRTFB	Major Range Test and Facility Base

Acronym	Definition
NAVAIR	Naval Air Systems Command
NAVSEA	Naval Sea Systems Command
NRL	Naval Research Laboratory
ONI	Office of Naval Intelligence
ONR	Office of Naval Research
OT	Operational Test
PCPAD	Planning & Direction, Collection, Processing & Exploitation, Analysis & Production, and Dissemination
RDT&E	Research, Development, Test & Evaluation
RTA	Runtime Assurance
SEI	Software Engineering Institute
SOCOM	United States Special Operations Command
SPAWAR	Space and Naval Warfare Systems Command
SpeAR	Specification and Analysis of Requirements
STANAG	Standardization Agreement
STAT COE	Scientific Test and Analysis Techniques Test & Evaluation Center of Excellence
SUAS	Small Unmanned Aircraft System
SUT	System under Test
T&E	Test and Evaluation
TARDEC	Tank Automotive Research, Development and Engineering Center
TEMP	Test and Evaluation Master Plan
TEP	Test and Evaluation Plan
TRL	Technology Readiness Level
TRMC	Test Resource Management Center
TTAs	Test Technology Areas
U.S. Army ATC	U.S. Army Aberdeen Test Center
U.S. Army RSJPO	U.S. Army Robotic Systems Joint Project Office
UAS	Unmanned Aerial System
UAST	Unmanned and Autonomous Systems Test
UAV	Unmanned Aerial Vehicle
UCS	UAS Control Segment Architecture
USD(AT&L)	Under Secretary of Defense for Acquisition, Technology and Logistics

Appendix C: Contributing Authors

Name	Function	Organization

Alley, Jim	Charter member	MCIA - Marine Corps Intelligence Activity
Clark, Matthew	Charter member	AFRL
Deal, Paul	Charter member	ONI - Office of Naval Intelligence
DePriest, Jeffrey	Charter member	DTRA/J9-CXSD
Hansen, Eric C CTR	Charter member	TRMC
Heitmeyer, Connie	Charter member	NRL
Nameth, Richard LTC	Charter member	DTRA/J9-CXWA
Steinberg, Marc	Charter member	ONR
Turner, Craig	Charter member	Army Evaluation Center
Young, Stuart	Charter member	ARL
Ahner, Darryl	Contributor	OSD STAT in T&E Center of Excellence
Alonzo Kelly	Contributor	CMU
Barry Bodt	Contributor	ARL
Friesen, Pete	Contributor	TASC
Jim Horris	Contributor	Johns Hopkins Univ APL
Jonathan Hoffman	Contributor	AFRL
Kerianne Gross	Contributor	AFRL
Laura Humphrey	Contributor	AFRL
Marshal Childers	Contributor	ARL
Michael Corey	Contributor	AFRL
Mike Wagner	Contributor	CMU
Reed, Mike	Contributor	ARA
Ryan Turner	Contributor	AFRL
Sandberg, Bill	Contributor	TASC
Schmidt, Mark	Contributor	ARA
Sean Regisford	Contributor	AFRL
Stanley Bak	Contributor	AFRL
Terwelp, Chris	Contributor	ARA
Bornstein, Jonathan	Coordination	ARL
Dowling, Steve	Coordination	DTRA/J9-CX
Hudas, Gregory	Coordination	TARDEC
Kearns, Kristen	Coordination	AFRL 711 HPW/RH
Kim, Dai H	Coordination	OSD OUSD ATL
Overholt, Jim	Coordination	AFRL 711 HPW/RH
Schuette, Lawrence	Coordination	ONR
Schultz, Alan	Coordination	NRL

Bamberger, Robert	Participant	Johns Hopkins Univ APL
Baran, David	Participant	US Army AMRDEC
Boydston, Alex	Participant	US Army AMRDEC
Chalmers, Robert	Participant	Johns Hopkins Univ APL
Condon, Sara	Participant	AMRDEC
Grabowski, Bob	Participant	MITRE - TRMC Support
Gross, Kerianne	Participant	AFRL/RQQD
Hinds, Bruce	Participant	DTRA/J9-CXT
Hinton, Mark	Participant	Johns Hopkins Univ APL
Hulbert, Brian	Participant	LinQuest
Laurri, Mark	Participant	OSD OUSD ATL
Lewis, Bruce	Participant	AMRDEC
McLoughlin, Mike	Participant	Johns Hopkins Univ APL
O'Donnell, Chris	Participant	OUSD AT&L
Panei, Vernon	Participant	TRMC
Rajkowski, Jessica	Participant	MITRE - TRMC Support
Reese, Shad	Participant	JGRE
Riddle, Stephanie	Participant	TRMC
Roberts, Rusty	Participant	GTRI
Rotner	Participant	MITRE - TRMC Support
Scheidt, David	Participant	Johns Hopkins Univ APL
Strausberger, Donald	Participant	GTRI
Tunstel, Edward	Participant	Johns Hopkins Univ APL
Whitehead, Steven	Participant	NAVSEA
Williams, Randy (Lewis)	Participant	BAH - JGRE Support
Young, Reed	Participant	Johns Hopkins Univ APL
Zeher, Michael	Participant	Johns Hopkins Univ APL